Object Extraction in Single Concept Video Frames Based on Visual and Motion Saliency Detection

Priya Devi S[1] , Surendheran A R[2]

[1]*Department of Computer Science and Engineering, KSR College of Engineering, Tamil Nadu, India;*
[2]*Department of Computer Science and Engineering, KSR College of Engineering, Tamil Nadu, India.*

Abstract

This paper presents automatic object instance extraction from single-concept video sequence frames. This object instance extraction fully their visual and motion saliency based. Visual and motion saliency information are extracted from their input video sequence frames. That saliency features are integrated by CRF method. It will be able to produce satisfactory results.

1. Introduction

A person can able to view the object attention in a video, although that object is obtainable in an unidentified or messy background or still never has been seen than earlier. The composite abilities displayed by person mind, this course of action to take synchronized extraction from a video. Devoid of some earlier information about the object attention or exercise data are still extremely difficult. As a result, if one must to be design an algorithm to automatically extract that object instance from the video frame, quit a few responsibilities are must be deal.

1) Undefined object type with an amount of undefined object case in video frames.
2) Difficult or else unpredicted movement of objects appropriate to expressed element or random poses.
3) Uncertain form among foreground and background areas appropriate to parallel color, low disparity, deficient illumination, etc. conditions.

Conversely, if anyone can extort delegate information from foreground either background regions from a video. The extracted information can be exploited between regions, and therefore the assignment of object instance extraction can be addressed.

In this paper, propose object instance extraction framework, which exploit mutually visual and motion saliency information crosswise video frames. The experimental saliency information permit us some visual and motion cues for knowledge foreground and background models and a CRF is applied to involuntarily resolve the label of each pixel based on the experiential representations. It will be able to produce satisfactory results. It will focus single concept videos with multiple object instances with pose, scale, etc. variations.

This paper is organized as follows: Section 2 introduce related works on object instance extraction and highlights the hand-outs of our method. Our proposed constructions are presented in sections 3 and 4. Finally section 5 concludes this paper.

2. Related Work

In universal, one can address VOE troubles using supervised or unsupervised approaches. Supervised approaches need previous information on the object of attention and require that collect exercise data in advance for designing the related VOE algorithms.

Unsupervised approaches do not train any explicit object detectors or classifiers in advance. For the videos captured by a static camera, extraction of object instances can be treated as a background subtraction problem. In other words, objects can be perceived basically by subtracting the current frame from a video sequence [1], [2]. However, if the background is every time varying or is occluded by object instances, background modeling turn into exceptionally demanding assignment. For such cases, investigators usually mean at information the background model from the input video sequence frame, and the foreground object instance are considered as outliers to be perceive. For instance, an auto regression moving average model (ARMA) that approximate the natural exterior of self-motivated surface and area was proposed in [3], and it mostly compact with situation in which the background consists of normal view similar to sea waves or trees. Sun *et al*. [4] develop color gradients of the background to find out the limitations of the object instance. Some unsupervised advance aim at

survey features related with the foreground object for VOE. For example, graph-based methods [5], [6] identify the foreground object regions by reduce the cost between neighboring unknown nodes/pixels information. Additional exclusively, one can segment the object instance by separating a graph addicted to displace pieces whose entire energy is reduced devoid of using any preparation information. Although moving consequences information in [5], [6], these approaches usually guess that the background/camera motion is central across video frames. For universal videos imprison by freely moving cameras, these technique may not simplify well. Different from graph-based methods, Leordeanu and Collins [7] proposed to monitor the co-occurrences of object features to identify the object instance in an unsupervised setting. Although hopeful consequences below pose, scale, occlusion, etc. variations were statement, their approach was only able to deal with rigid objects (like cars).

Since Itti *et al*. [8] first derived the visual saliency of a single representation, several mechanism have been planned to extract the saliency information of images for the tasks of compression, classification, or segmentation. For exemplar, Harding and Robertson [9] exhibit that the visual saliency can be develop to recover image density ratio by merge SURF features and task-dependent previous information. Unlike density or categorization problems which might make the most of task or object category information for receive the related saliency, universal saliency detection or image segmentation tasks are resolve in an unsupervised setting. For example, standard spectrum analysis, Hou and Zhang [10] utilized the spectral residual as saliency information, while Guo *et al*. [11] superior stage of the range simultaneously with Quaternion Fourier Transform for saliency detection. Liu *et al*. [12] measured dissimilarity information and color histogram of dissimilar image areas in various scale to identify neighborhood and large-scale image saliency. Achanta and Süsstrunk [13] calculate the saliency by enchanting symmetric neighboring pixels into consideration and averaging the color dissimilarity among pixels within every area. Goferman *et al*. [14] applied multi-scale piece and considered both color dissimilarity and positions among dissimilar space. Zhai and Shah [15] create spatial and sequential saliency maps by using a spatiotemporal consideration model. Supported on neighborhood image contrast, Ma and Zhang [16] resolute the salient areas by fuzzy growing which extracts regions or objects of interest when structure the saliency map. Recently, Wang *et al*. [17] project a genetic stimulated approach and consequential visual saliency based on site entropy velocity for saliency detection. Nevertheless, discovering visual saliency in images or video frames would present hopeful consequences and infer the region of the foreground object instance. Conversely, since real-world videos might encounter low contrast or unsatisfactory lighting, etc. problems, one might not be able to achieve advantageous visual saliency maps for recognize foreground object instances. As a result, one cannot simply apply visual saliency methods for segmenting object instance in real world videos.

3. OBJECT INSTANCE MODELING AND SEGMENTATION

A large amount of obtainable unsupervised approaches presume the objects are outliers in stipulations of the observed motion information. In this paper work is saliency-based object extraction framework which learns saliency information in both spatial (visual) and temporal (motion) domains. By move on conditional random fields (CRF), the integration of the resulting features can robotically categorize the object instance without the need to treat either foreground or background as outliers. Fig. 1 shows the outline of our framework.

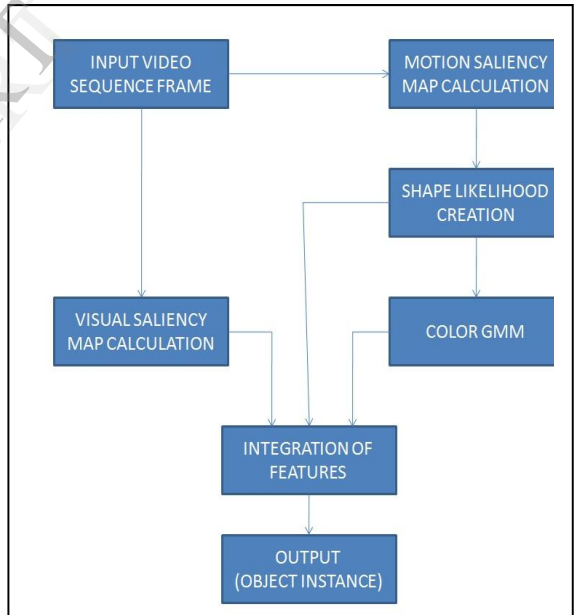


Fig. 1 . Outline of our framework

3.1 Visual Saliency Calculation

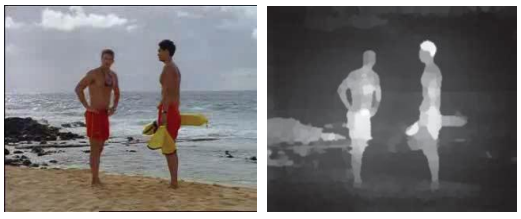
To extract visual saliency of each frame, we perform image segmentation on each video frame and extort color and contrast information. Turbo pixels proposed by [18] for, and the resulting image segments (superpixels) are applied to perform saliency detection. For the *k*th superpixel r_k, we calculate its saliency score $S(r_k)$ as follows:

$$S(r_k) = \sum_{r_k \neq r_i} \exp(D_s(r_k, r_i)/\sigma_s^2)\omega(r_i)D_r(r_k, r_i)$$
$$\approx \sum \exp(D_s(r_k, r_i)/\sigma_s^2)D_r(r_k, r_i) \quad (1)$$

Where D_s is the Euclidean distance between the centric of r_k and that of its surrounding superpixels r_i, while σ_s controls the width of the kernel. The parameter $\omega(r_i)$ is the weight of the neighbor superpixel r_i, which is proportional to the number of pixels in r_i. Compared to [19], $\omega(r_i)$ can be treated as a constant for all superpixels. The last term $D_r(r_k, r_i)$ measures the color variation between r_k and r_i, which is also in terms of Euclidean distance. We regard the pixel i as a salient point if its saliency score satisfies $S(i) > 0.8 * \max(S)$, and the collection of the consequential salient pixels will be considered as a salient point set. Since image pixels which are nearer to this salient point set should be visually more important than those which are beyond away, we additional refine the saliency $\hat{S}(i)$ for each pixel i as follows:

$$\hat{S}(i) = S(i) * (1 - \text{dist}(i)/\text{dist}_{\max}) \quad (2)$$

$S(i)$ is the original saliency score derived by (1), and $\text{dist}(i)$ measures the adjacent Euclidian distance to the salient point set. We note that $\text{dist}_{\max}$ in (2) is determined as the most detachment from a pixel of attention to its adjacent salient point within an image, thus it is an image-dependent constant. An example of visual saliency calculation is shown in fig.2.



(a) (b)

Fig. 2. (a)Original video frame. (b) visual saliency map

3.2 Calculation For Motion Saliency Induced Cues

3.2.1 Motion Saliency Calculation:

In our work, each moving part of a foreground object is implicit. To discover the moving parts and their equivalent pixels, we perform dense optical-flow forward and backward dissemination [21] at every frame. A moving pixel qt at frame t is determined by:

$$q_t = \widehat{q}_{t,t-1} \bigcap \widehat{q}_{t,t+1} \quad (3)$$

$\hat{q}$ denotes the pixel pair detected by forward or backward optical flow dissemination. Only if a pixel is recognized by the optical-flow path in both instructions, we will denote it as a pixel of a moving object. After determining the areas stimulate by the moving object (or its parts), we will extract the related shape and color information from these regions.

(a) (b)

Fig. 3. (a)Original video frame. (b) Motion saliency map

3.2.1 Shape Likelihood Creation:

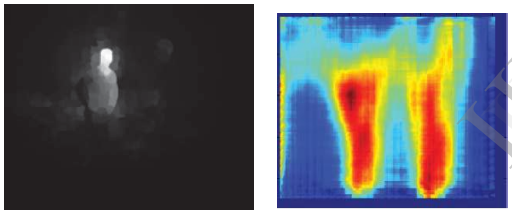
Since we assume each moving part of an object forms a complete sampling of the entire object, part-based shape information induced by above motion cues can be advanced to characterize the foreground object. To describe each moving part, we apply the histogram of oriented gradients (HOG) features. We first divide each frame into disjoint 8×8 pixel grids, and we compute HOG descriptors for each region (patch) of 4×4 = 16 grids. To capture range invariant outline in sequence, additionally we reduce the resolution of every frame and do again the above procedure. We reminder [22] also used a parallel setting to extract their HOG descriptors. Since the use of sparse illustrations has been made known to be extremely effective in many computer vision tasks [23], once the HOG descriptors of the moving foreground regions are extracted, we learn an over-complete codebook and find out the related sparse illustration of each HOG. Now, for a total of N HOG descriptors calculated for the above motion-salient areas $\{h_n, n = 1, 2,\ldots,N\}$ in a p- dimensional space, we construct an over-complete dictionary $D^{p\times K}$ which contain K basis vectors, and we conclude the equivalent sparse coefficient α_n of each HOG descriptor. Simply speaking, the sparse coding problem can be formulated as:

$$\min_D \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}||\mathbf{h}_n - \mathbf{D}\alpha_n||_2^2 + \lambda||\alpha_n||_1 \quad (4)$$

where λ balances the scarcity of α_n and the l_2-norm reconstruction error. We use the software developed by [24] to solve the above problem. We further calculate the mask M for each codeword by averaging the moving regions with coefficient α_n. Behind attain the dictionary and the masks to represent the shape of objet instance, we use them to encode all image patches at each frame. This is to recover non-moving regions of the foreground object instance which does not have important motion and thus cannot be perceive by motion cues. For each image patch, we derive its sparse coefficient vector α, and each access of this vector point out the part of each shape codeword. Likewise, we use the related masks and their influence coefficients to evaluate the concluding mask for each image patch. The reconstruction image using foreground shape information is then formulated as:

$$\widehat{X}_t^{\mathcal{S}} = \sum_{n \in I_t} \sum_{k=1}^{K} (\alpha_{n,k} \cdot M_k) \tag{5}$$

Figure 4 shows an example of the reconstruction of a video frame using shape information of the foreground object (induced by motion cues only).



(a) (b)

Fig. 4. (a)Motion saliency map. (b) Shape likelihood

3.2.3 Learning Of Color Cues

Foreground object regions which are not salient in terms of visual or motion form will be consider as background, and the consequential color models will not be of satisfactory discerning capability. In our work, we utilize the shape likelihood _ *XS t* gained from the prior step, and we threshold this likelihood by 0.5 to determine the candidate foreground (*FS*shape) and background (*BS*shape) areas. In other words, we consider color information of pixels in *FS*shape for calculating the foreground color GMM, and those in *BS*shape for deriving the background color GMM. At last, we combine both foreground and background color models with visual saliency and shape likelihood into a united framework for object extraction.

4 CONDITIONAL RANDOM FIED FOR OBJECT INSTANCE EXTRACTION

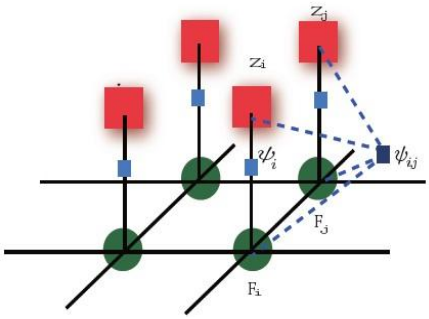


Fig. 5. CRF for object segmentation

Utilizing an undirected graph, conditional random field (CRF) [20] is an authoritative method to approximate the structural information of a set of variables with the related interpretation. For video foreground object segmentation, CRF has been applied to calculate the label of each experimental pixel in an image *I* [5], [6]. As illustrated in Fig. 5, pixel *i* in a video frame is related with remark z_i, while the unknown node F_i point out its related label. In this framework, the label F_i is calculated by the observation z_i , while the spatial coherence between this output and neighboring remarks z_j and labels F_j are concurrently taken into consideration. Therefore, predicting the label of an observation node is equivalent to exploit the following posterior probability function

$$p(F|I,\psi) \propto \exp\left\{-\left(\sum_{i \in I}(\psi_i) + \sum_{i \in I, j \in \text{Neighbor}}(\psi_{i,j})\right)\right\} \tag{6}$$

where ψ_i is the unary term which infers the likelihood of F_i with observation z_i . $\psi_{i,j}$ is the pair wise term relating the correlation among neighboring pixels z_i and z_j, and that among their predicted output labels F_i and F_j . Note that the observation *z* can be represented by a particular feature or a grouping of various types of features. To solve a CRF optimization problem, one can convert the above problem into an energy minimization task, and the object energy function *E* of (6) can be derived as

$$\begin{aligned} E &= -\log(p) \\ &= \sum_{i \in I}(\psi_i) + \sum_{\substack{i \in I \\ j \in \text{Neighbor}}}(\psi_{i,j}) \\ &= E_{\text{unary}} + E_{\text{pairwise}}. \end{aligned} \tag{7}$$

In our proposed framework, we define the shape energy function E^S in terms of shape likelihood X_t^S (derived by (5)) as one of the unary terms

$$E^{\mathcal{S}} = -w^s \log(\widehat{X}_t^{\mathcal{S}}). \tag{8}$$

In addition to shape information, we need incorporate visual saliency and color cues into the introduced CRF framework. As discussed earlier, we derive foreground and background color models for object extraction, and thus the unary term E^C describing color information is defined as follows:

$$E^{\mathcal{C}} = w^c (E^{\mathcal{CF}} - E^{\mathcal{CB}}). \tag{9}$$

Note that the foreground and background color GMM models $G^C{}_f$ and $G^C{}_b$ are utilized to originate the related energy terms E^{CF} and E^{CB}, which are calculated as

$$\begin{cases} E^{\mathcal{CF}} = -\log\left(\sum_{i\in I} G^{C_f}(i)\right) \\ E^{\mathcal{CB}} = -\log\left(\sum_{i\in I} G^{C_b}(i)\right). \end{cases}$$

As for the visual saliency cue at frame t, we convert the visual saliency score $\hat{}St$ derived in (2) into the following energy term E^V:

$$E^{\mathcal{V}} = -w^v \log(\widehat{S}_t). \tag{10}$$

We note that in the above equations, parameters w^s, w^c, and w^v are the weights for shape, color, and visual saliency cues, correspondingly. These weights organize the contributions of the related energy terms of the CRF model for performing VOE. It is also worth noting that, Liu and Gleicher [5] only considers the construction of foreground color models for VOE. As verified by [6], it can be concluded that the disregard of background color models would limit the performance of object extraction, since the only use of foreground color model might not be enough for characteristic between foreground and background regions. In the proposed object extraction framework, we now utilize multiple types of visual and motion salient features for object extraction.

5. CONCLUSION

This paper will utilize visual and motion saliency induced features, based on that features we can able to extract the object instances in video sequence frame. We advanced CRF model, that model will integrate the above features. A major advantage of our method is that object instance extraction doesn't require any preceding knowledge about the object and also doesn't require any user interaction during the segmentation process.

6 . REFERENCES

[1] T. Bouwmans, F. E. Baf, and B. Vachon, "Background modeling using mixture of Gaussians for foreground detection—A survey," *Recent Patents Comput. Sci.*, vol. 3, no. 3, pp. 219–237, 2008.

[2] F.-C. Cheng, S.-C. Huang, and S.-J. Ruan, "Advanced background subtraction approach using Laplacian istribution model," in *Proc. IEEEInt. Conf. Multimedia Expo*, Jul. 2010, pp. 754–759.

[3] J. Zhong and S. Sclaroff, "Segmenting foreground objects from a dynamic textured background via a robust Kalman filter," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, vol. 1. Oct. 2003, pp. 44–50.

[4] J. Sun, W. Zhang, X. Tang, and H.-Y. Shum, "Background cut," in *Proc.9th Eur. Conf. Comput. Vis.*, 2006, pp. 628–641.

[5] F. Liu and M. Gleicher, "Learning color and locality cues for moving object detection and segmentation," in *Proc. IEEE Conf. Comput. Vis Pattern Recognit.*, Jun. 2009, pp. 320–327.

[6] K.-C. Lien and Y.-C. F. Wang, "Automatic object extraction in singleconcept videos," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2011, pp. 1–6.

[7] M. Leordeanu and R. Collins, "Unsupervised learning of object features from video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 1142–1149..

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[9] P. Harding and N. M. Robertson, "Visual saliency from image features with application to compression," *Cognit. Comput.*, vol. 5, no. 1, pp. 76–98, 2012.

[10] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[11] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[12] T. Liu, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*,Jun. 2007, pp. 1–8.

[13] R. Achanta and S. Süsstrunk, "Saliency detection using maximum symmetric surround," in *Proc. IEEE Int. onf. Image Process.*, Sep. 2010, pp. 2653–2656.

[14] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2376–2383.

[15] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM Int. Conf. Multimedia*, 2006, pp. 815–824.

[16] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 374–381.

[17] W. Wang, Y. Wang, Q. Huang, and W. Gao, "Measuring visual saliency by site entropy rate," in

Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 2368–2375.

[18] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, and S. J. Dickinson, "TurboPixels: Fast superpixels using geometric flows," *IEEE Trans.Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2290–2297, Dec. 2009.

[19] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu,"Global contrast based salient region detection," in *Proc. IEEE Conf.Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 409–416.

[20] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Data*. SanMateo, CA, USA: Morgan Kaufmann, 2001, pp. 282–289.

[21] M. Werlberger et al., "Anisotropic huber-L1 optical flow," in BMVC, 2009.

[22] P. Felzenszwalb, et al., "Object detection with discriminatively trained part based models," IEEE PAMI, 2010.

[23] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," IEEE Trans. Image Processing, 2006.

[24] J. Mairal et al., "Online learning for matrix factorization and sparse coding," JMLR, 2010.